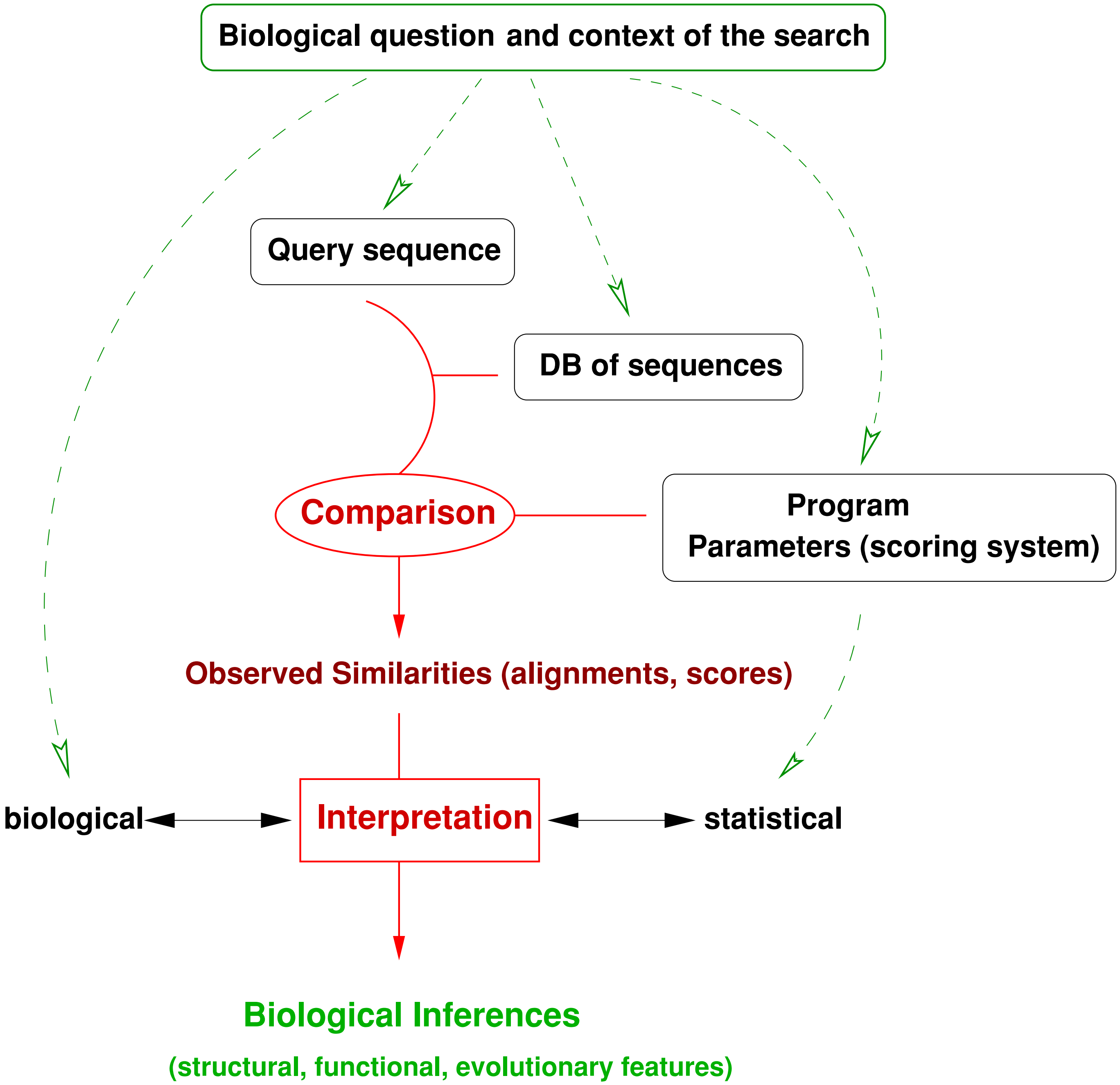
Part 2:
Similarity searching in sequence databases


Biological question and context of the search

Query sequence

DB of sequences

Comparison

Program
Parameters (scoring system)

Observed Similarities (alignments, scores)

biological

Interpretation

statistical

Biological Inferences

(structural, functional, evolutionary features)

Nucleotide Sequence Databases

The EMBL/GenBank/DDBJ international Nucleotide Sequence Database

Exhaustive

Redundant

Not curated

Heterogenous

Many specialized (organism specific) databases

Curated and annotated by experts of the organism

See Nucleic Acids Research, 1st of January, each year, and journals like Bioinformatics or Genome Research.

Protein Sequence Databases

Curated and Annotated Protein Resources

PIR

SwissProt

Translations of Nucleotide Databases

Genpept

TrEMBL : supplement to SwissProt

⟶ sptrnrdb, "SwissProt, TREMBL, Non Redundant DataBase"
= SwissProt + TrEMBL + TrEMBL_new

Composite Databases

NCBI–nrprot

OWL

Heuristic algorithms

They prune the search space by:

1: using fast approximate methods to select the sequences of the database that are likely to be similar to the query and to locate the similarity region inside them.

2: restricting the alignment process:

- only to the selected sequences
- only to some portions of the sequences

FASTA & BLAST story

1985 : FASTP (D. Lipman and W. Pearson)

Global gapped alignments

1988 : FASTA (W. Pearson and D. Lipman)

Local gapped alignments

1990 : BLAST1

(S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman)

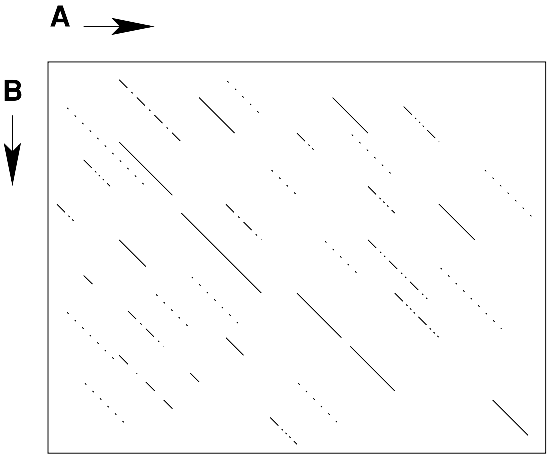
Local ungapped alignments

Gapped BLASTs :

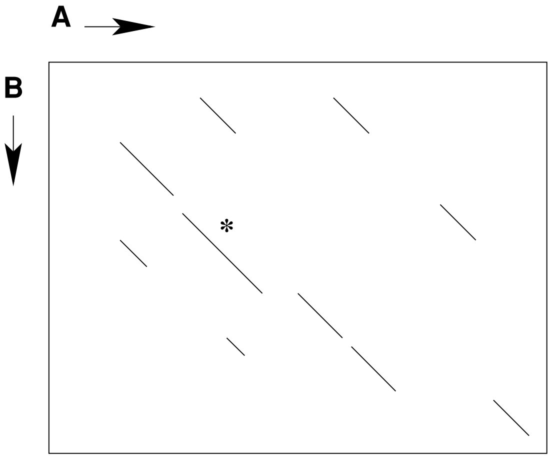
1996: WU–BLAST2 (W. Gish)

1997: NCBI–BLAST2 (and PSI–BLAST)
(S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang,
W. Miller and D. Lipman)

FASTA ALGORITHM

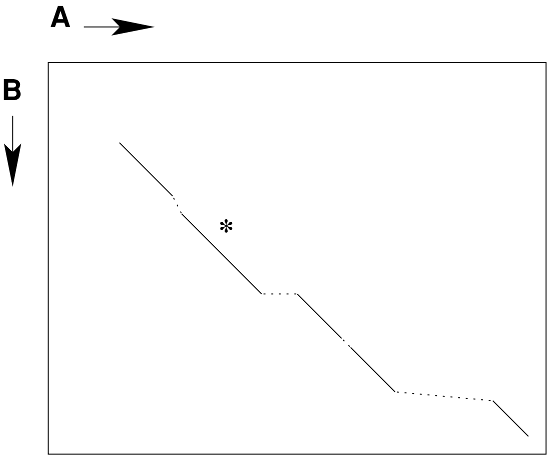


Identify all k–tuple matches



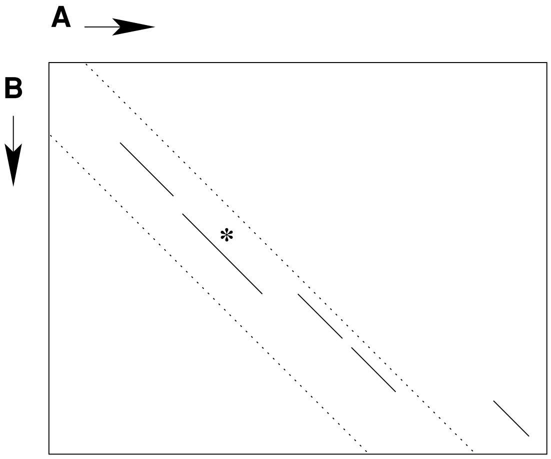
score the 10 best scoring regions using a scoring matrix

→ Init1 score



Apply joining procedure

→ Initn score



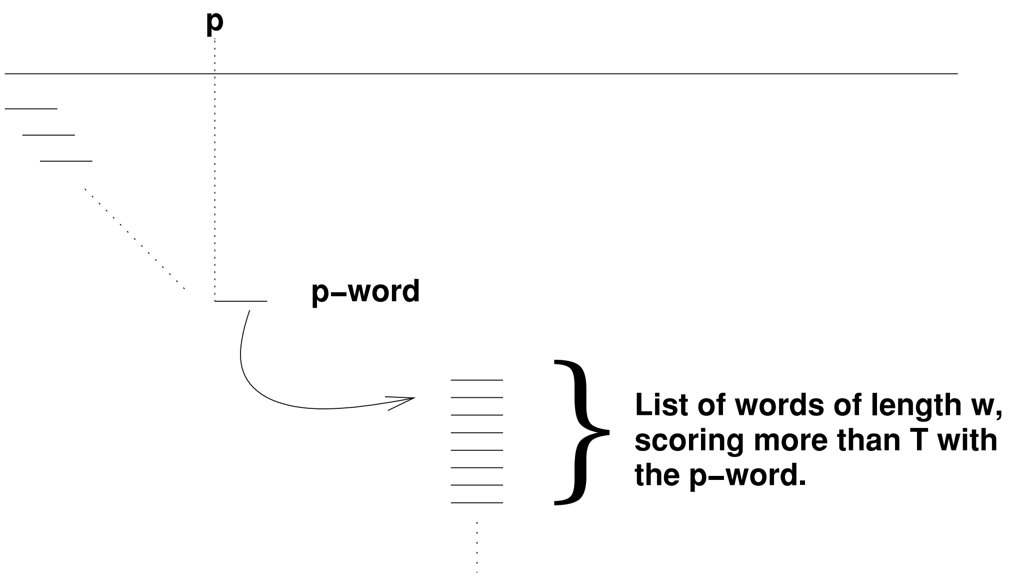
Apply limited DP

→ Opt score

BLAST1 ALGORITHM

First step:

For each position p of the query, find the list or words of length w scoring more than T when paired with the word starting at p:



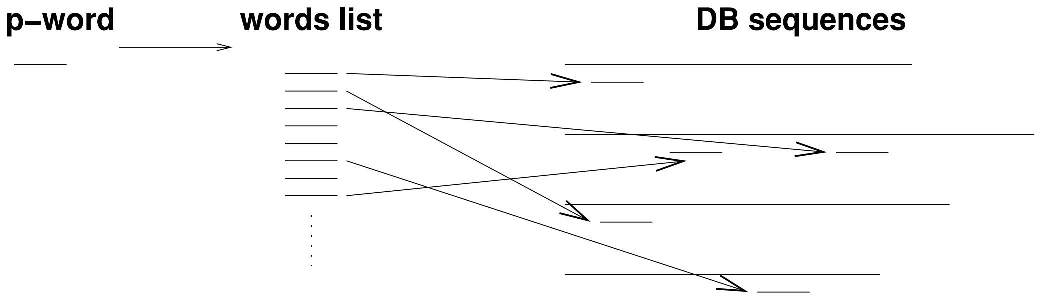
p

p–word

List of words of length w, scoring more than T with the p–word.

Second step:

For each words list, identify all exact matches with DB sequences:



p–word

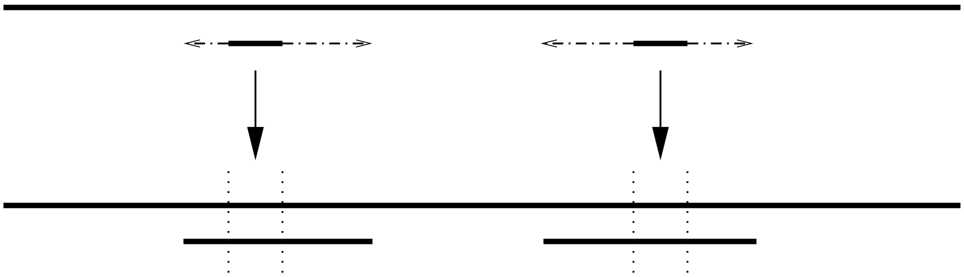
words list

DB sequences

BLAST1 ALGORITHM

Third step:

For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.



HSP = High Scoring Segment Pair

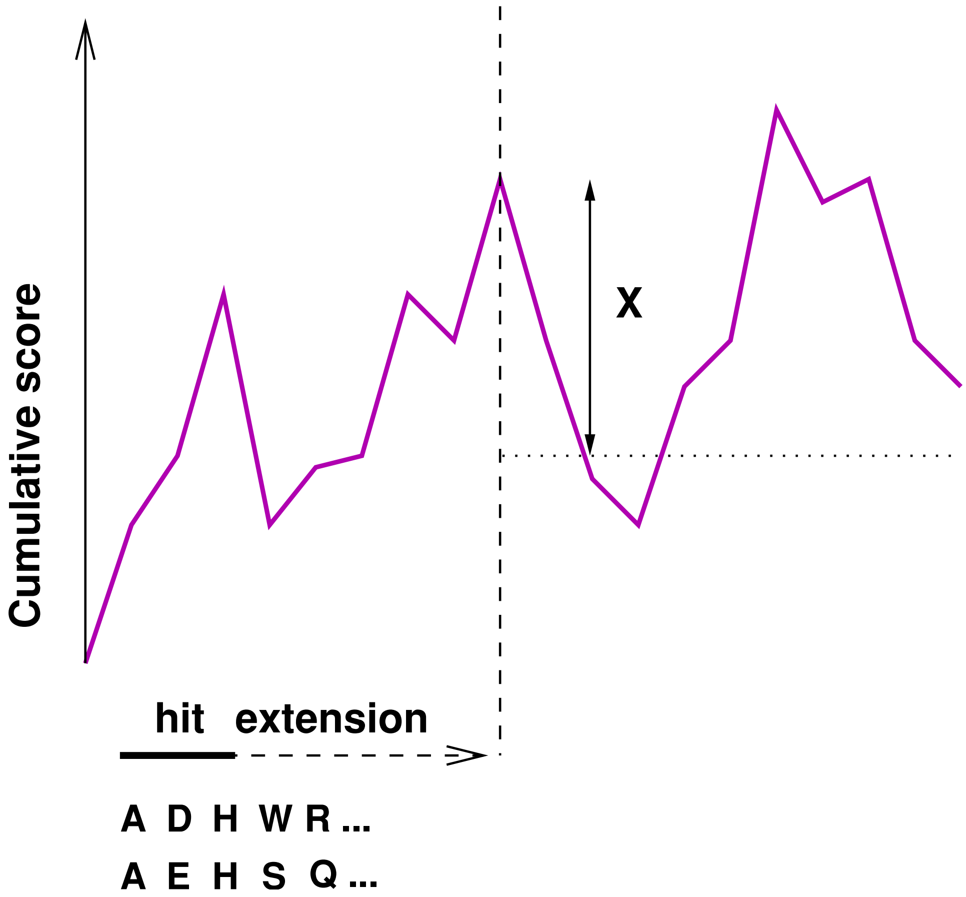
MSP = Maximal Segment Pair

Reports all HSPs having score S above a threshold, or equivalently, having E−value below a threshold.

E−value = the number of HSPs having score S (or higher) expected to occur only by chance.

Apply sum−statistics to evaluate the significance of a combination of HSPs involving the same DB sequence.

Ungapped extension of hits



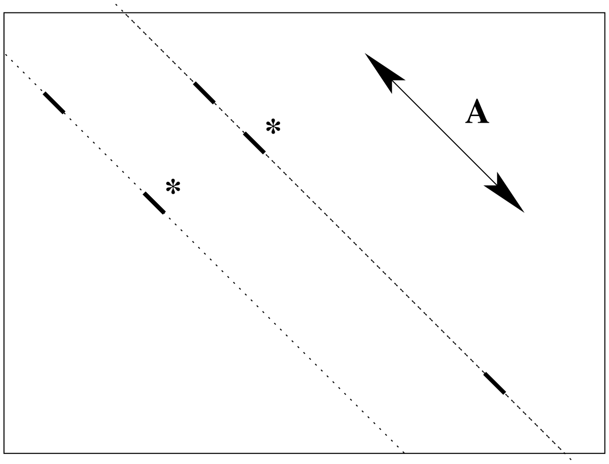
NCBI–BLAST2

The «two–hits» requirement

First step: as with BLAST1, generate lists of words scoring more than T with words of the query.

Second step: generation of hits: identify all word matches in DB sequences

Third step: extension of hits: requires a second hit on the same diagonal at a distance of less than A.



This step generates ungapped HSPs

Fourth step: gapped extension of HSPs having score above a threshold S_g

WU–BLAST2

First step: as with BLAST1, generate lists of words scoring more than T with words of the query.

Second step: generation of hits: identify all words matches with the DB sequences

Third step: ungapped extension of hits :

. default's behavior: «one–hit» requirement (as BLAST1)

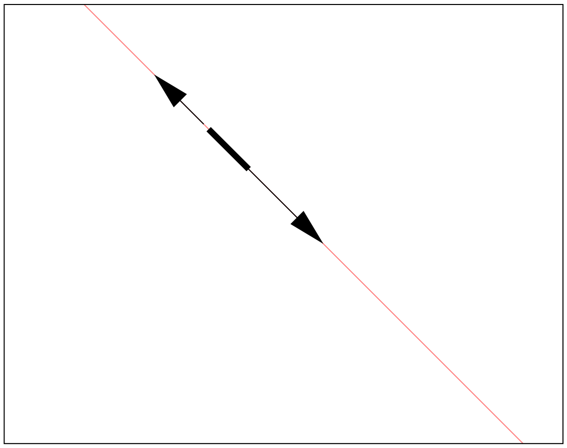
. «hitdist» option: «two–hits» requirement (as ncbi–BLAST2)

Fourth step: HSPs with score S above a threshold trigger gapped extensions

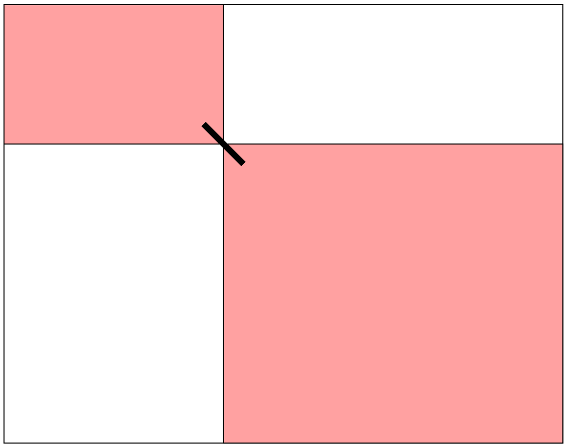
«nogap» option: fourth step is not performed

Evaluates the statistical signifance of multiple local alignments using «Sum statistics»

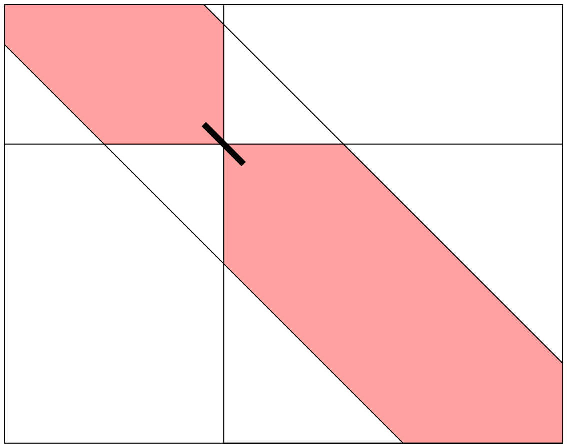
Ungapped and gapped extensions



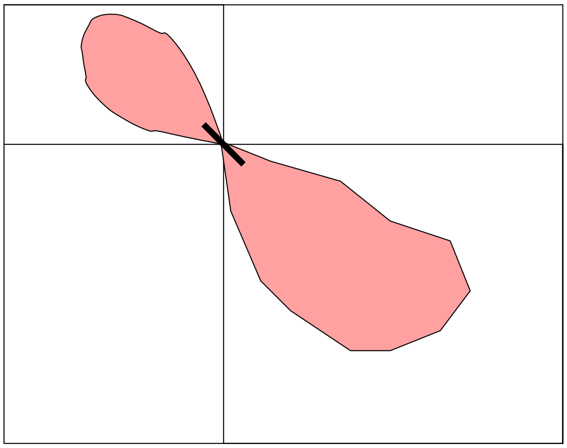
Ungapped extension



Gapped extension by full DP



Gapped extension by «banded DP»



Gapped extension by «score–limited DP»

Statistics of alignments scores

Question: Was is the probability of chance occurence of an alignment having score S or greater?

⟶ We need to know the random distribution of the scores, i.e. the distribution of alignment scores under a random model

Global alignments:

the distribution is not known

Local alignments without gaps:

theoretical work: Karlin−Altschul statistics

⟶ Extreme−value distribution

Local alignments with gaps:

empirical studies

⟶ Extreme−value distribution

Karlin–Altschul statistics

$\longrightarrow$ Apply to local ungapped alignments

Random Model:

– Random sequences:

Independent and identically distributed residues, taken with background probabilities p_i , p_j .

– Random variable:

S, score of the MSP (Maximal Segment Pair)

– Scoring system:

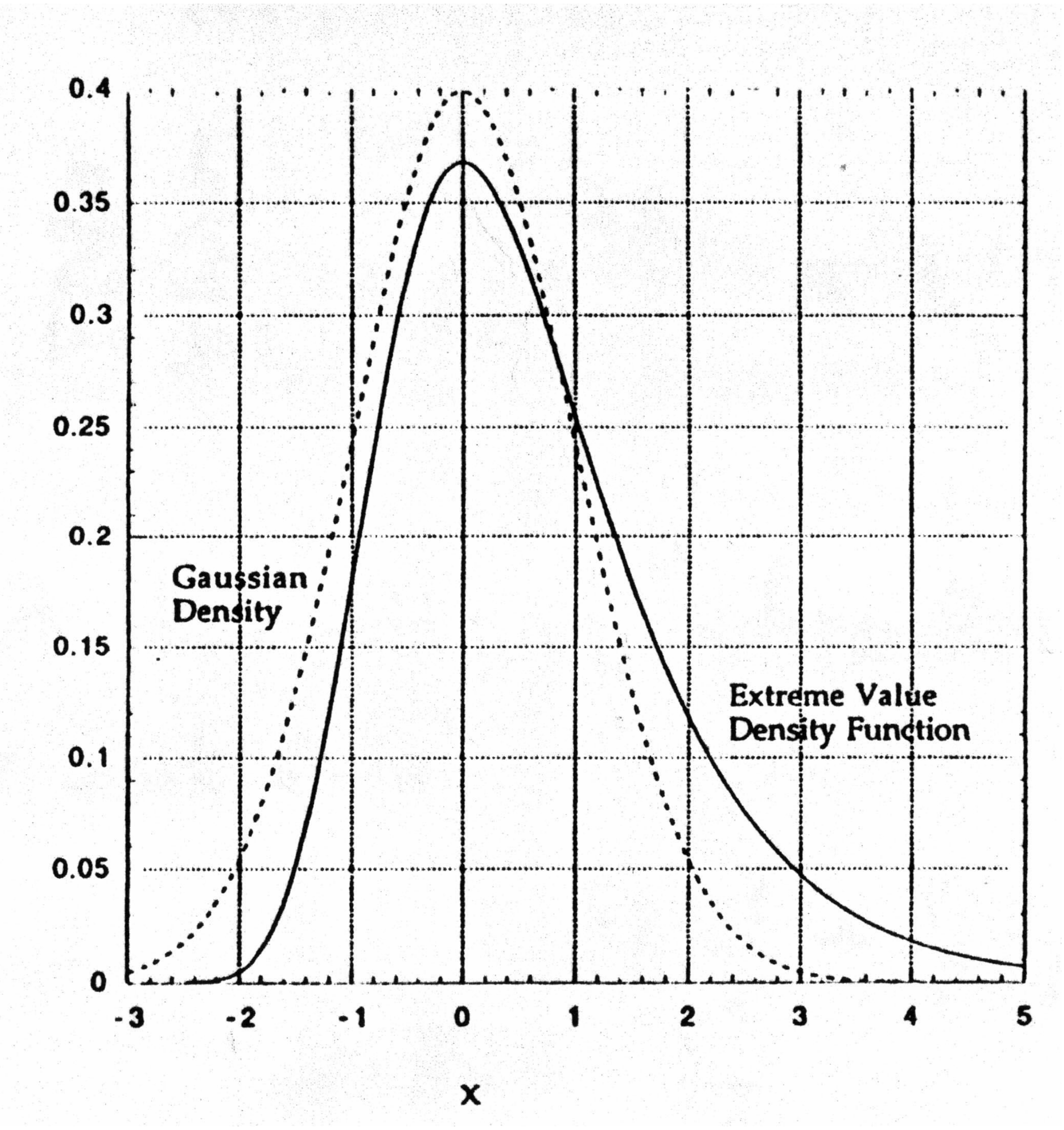
A set of similarity scores, $S_{i,j}$, such as:

- at least one of the scores $S_{i,j}$ is strictly positive
- the expected score for a random pair of residues has to be negative:

$$\sum_{i,j} p_i p_j S_{i,j} \leq 0$$

Under this random model and given that the lengths of the two sequences being compared are large, S follows an Extreme–Value distribution.

The Extreme Value Distribution



(from W. Gish, 1996)

```
searching /local/databases/fasta/sptrnrdb library

       opt      E()
< 20   994     0:=
  22     3     0:=          one = represents 1377 library sequences
  24    50     1:*
  26   165    17:*
  28   700   186:*
  30  2785  1129:*==
  32  6654  4364:===*=
  34 14518 11834:========*==
  36 24183 24303:=================*
  38 40186 40164:=============================*
  40 55669 56026:========================================*
  42 68512 68485:=================================================*
  44 79155 75545:======================================================*===
  46 82616 76945:=======================================================*====
  48 80086 73666:=====================================================*=====
  50 68245 67220:================================================*=
  52 59186 59098:==========================================*
  54 49603 50480:====================================*
  56 39874 42166:============================= *
  58 34221 34618:=========================*
  60 26309 28042:====================*
  62 21374 22482:================*
  64 14726 17879:=========== *
  66 11964 14131:========= *
  68  9241 11116:======= *
  70  6784  8711:===== *
  72  5428  6807:====*
  74  3915  5307:===*
  76  3000  4130:==*
  78  2308  3211:==*
  80  1702  2493:=*
  82  1376  1907:=*
  84   977  1511:=*
  86   631  1169:*
  88   516   904:*          inset = represents 7 library sequences
  90   400   700:*
  92   311   541:*         :=======================================
  94   208   419:*         :==============================                                   *
  96   184   324:*         :===========================                  *
  98   141   251:*         :=====================              *
 100    88   194:*         :=============              *
 102    95   150:*         :==============       *
 104    60   116:*         :=========       *
 106    45    90:*         :=======     *
 108    31    70:*         :=====    *
 110    31    54:*         :=====  *
 112    27    42:*         :==== *
 114    16    32:*         :=== *
 116     7    25:*         :=  *
 118     2    19:*         := *
>120   112    15:*         :==*=============
257666599 residues in 819414 sequences
 statistics extrapolated from 60000 to 819137 sequences
```

Karlin–Altschul statistics

p–value: **probability that there is at least one random MSP having score S or greater.**

$$p(score \geq S) = 1 - exp(-Kmne^{-\lambda S})$$

E–value: **expected number of random MSP having score S or greater.**

$$E(S) = Kmne^{-\lambda S}$$

Analytical formulas are available, enabling to calculate λ **and** K **from the parameters of the random model (i.e. background probabilities, similarity scores, lengths of the sequences)**

Normalized scores: $S' = \lambda S - \ln K$

Bit scores: $S' = \frac{\lambda S - \ln K}{\ln 2}$ $\quad E(S') = mn2^{-S'}$

Statistics of local gapped alignments

Empirically shown that they follow an extreme−value distribution.

Need of empirical simulations of the random distribution in order to calculate its parameters.

Blast2 (both of them):

artificial random sequences

Fasta:

uses results from the search: real unrelated sequences